

RESEARCH

Open Access



# DrugReAlign: a multisource prompt framework for drug repurposing based on large language models

Jinhang Wei<sup>1</sup>, Linlin Zhuo<sup>1\*</sup>, Xiangzheng Fu<sup>2\*</sup>, XiangXiang Zeng<sup>3</sup>, Li Wang<sup>4</sup>, Quan Zou<sup>5</sup> and Dongsheng Cao<sup>6\*</sup>

## Abstract

Drug repurposing is a promising approach in the field of drug discovery owing to its efficiency and cost-effectiveness. Most current drug repurposing models rely on specific datasets for training, which limits their predictive accuracy and scope. The number of both market-approved and experimental drugs is vast, forming an extensive molecular space. Due to limitations in parameter size and data volume, traditional drug-target interaction (DTI) prediction models struggle to generalize well within such a broad space. In contrast, large language models (LLMs), with their vast parameter sizes and extensive training data, demonstrate certain advantages in drug repurposing tasks. In our research, we introduce a novel drug repurposing framework, DrugReAlign, based on LLMs and multi-source prompt techniques, designed to fully exploit the potential of existing drugs efficiently. Leveraging LLMs, the DrugReAlign framework acquires general knowledge about targets and drugs from extensive human knowledge bases, overcoming the data availability limitations of traditional approaches. Furthermore, we collected target summaries and target-drug space interaction data from databases as multi-source prompts, substantially improving LLM performance in drug repurposing. We validated the efficiency and reliability of the proposed framework through molecular docking and DTI datasets. Significantly, our findings suggest a direct correlation between the accuracy of LLMs' target analysis and the quality of prediction outcomes. These findings signify that the proposed framework holds the promise of inaugurating a new paradigm in drug repurposing.

**Keywords** Drug repositioning, Large Language Model, Drug-target interactions, Molecular docking

## Background

Drug development is a complex, interdisciplinary endeavor that encompasses a sequence of intricate, time-consuming, and costly steps, including target identification, drug screening, molecular optimization, clinical trials, and regulatory approval. The journey from clinical evaluation to market launch for a new drug often spans over a decade and incurs a cost of approximately one billion US dollars [1]. Moreover, the success rate for new drug development is exceedingly low; even potential compounds that enter the clinical phase have only about an 8% chance of receiving regulatory approval [2]. Drug repositioning, an attractive strategy within drug development, aims to repurpose existing drugs or compounds for new diseases beyond their original indications. This approach significantly simplifies the drug development

\*Correspondence:  
Linlin Zhuo  
zhuoninnin@163.com  
Xiangzheng Fu  
fxzheng@hkbu.edu.hk  
Dongsheng Cao  
oriental-cds@163.com

<sup>1</sup> School of Data Science and Artificial Intelligence, Wenzhou University of Technology, Wenzhou 325027, China

<sup>2</sup> School of Chinese Medicine, Hong Kong Baptist University, Hong Kong 519087, China

<sup>3</sup> College of Computer Science and Electronic Engineering, Hunan University, Changsha 410012, China

<sup>4</sup> Department of Computer Science, University of Tsukuba, Tsukuba 3058577, Japan

<sup>5</sup> Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 611730, China

<sup>6</sup> Central South University, Hunan University, Changsha 410083, China



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

process while maximizing the potential of existing drugs. For instance, aspirin was initially developed as an analgesic and antipyretic. Thanks to its antiplatelet aggregation properties, the drug helps reduce thrombus formation, thereby lowering the risk of cardiovascular events and has been applied in the prevention of heart disease and stroke [3]. Sildenafil, originally developed for the treatment of angina pectoris, was later discovered to have significant therapeutic effects on erectile dysfunction (ED) and thus was repurposed as a medication for treating ED [4].

Traditional drug repositioning methods necessitate the analysis of extensive medical literature and clinical data, collaborating across disciplines such as pharmacology and chemistry to uncover potential diseases that existing drugs can treat [5]. This process is both labor-intensive and talent-intensive, resulting in relatively low efficiency [6]. In recent years, with the rapid development of bioinformatics, computational approaches, particularly those based on deep learning, have achieved remarkable progress in the field of drug repositioning [7–9]. These methods have the ability to automatically extract intricate features from both drugs and targets, significantly improving the accuracy and interpretability of predictive models. This has greatly advanced the potential for discovering new drug applications and accelerated the development of drug repositioning. Moreover, ongoing innovations in computational methods continue to expand the capabilities of drug repositioning. Approaches that blend deep learning with advanced techniques like fuzzy logic and graph clustering are helping to address challenges related to data complexity and uncertainty [10]. While these approaches often utilize specific data sources, they offer valuable insights that can be further enhanced to achieve broader generalizability in future research.

In recent years, Natural Language Processing (NLP) has advanced significantly. Models such as ChatGPT, which are examples of LLMs, have been trained on extensive human knowledge databases. They exhibit exceptional performance in tasks like question answering, translation, and sentiment analysis [11]. Unlike traditional rule-based or knowledge-based models, LLMs autonomously learn language rules and patterns, facilitating a profound grasp of underlying knowledge [12]. Their interdisciplinary learning capability heralds a potential shift in scientific research domains, including drug discovery. LLMs efficiently deliver insights on pertinent topics through streamlined dialogue, considerably reducing the time and accelerating scientific progress. Furthermore, their adeptness at processing and analyzing data from diverse sources, converting various data forms into textual descriptions, proves invaluable [12]. This ability is especially crucial in fields dealing with

heterogeneous, interdisciplinary data, such as genomics [13], protein engineering [14], material science [15] and drug development [16, 17]. Domain-specific models such as PMC-LLaMA and Me-LLaMA, fine-tuned on biomedical corpora, demonstrate even greater potential in applications related to life sciences [18, 19]. By harnessing extensive multi-source information, LLMs quicken and enrich the data analysis process, thereby augmenting the efficiency and comprehensiveness of scientific research. These developments suggest LLMs could revolutionize the scientific research landscape in the foreseeable future.

The development of LLMs has infused new vitality into the field of drug repositioning. The training data of LLMs includes an extensive array of medical literature related to target drugs, covering not only basic information about the drugs, their pharmacological actions, and clinical trial outcomes but also detailed knowledge about the relationships between drugs and diseases, targets, and biological pathways. By integrating these vast and diverse data sources into their knowledge base, LLMs can provide comprehensive and multidimensional reference information for drug repositioning tasks, enhancing their precision and interpretability. Undeniably, LLMs also face issues such as hallucinations, dissemination of false information, and misunderstandings, which could lead to the provision of misleading answers [20]. At its core, the primary issue is that most existing methods rely on prompts from a single source, leading to inevitable statistical biases.

To address these challenges, we introduced a drug repositioning framework leveraging LLMs and multi-source prompting techniques to fully harness the potential of existing drugs. This study sourced extensive target protein information from the RCSB database and applied a structured method to extract vital prompt details such as PDB ID and PDB Name. We also detailed the binding pockets of small molecule ligand-target complexes from their PDB files, incorporating atomic coordinates and spatial interaction data. Using pocket analysis tools, we identified key characteristics like hydrophobicity and hydrogen bonds. Following data cleaning, prompt information was aggregated from 1,273 PDB structures. These structures, combined with spatial interaction cues, were analyzed using various LLMs (GPT-3.5 [21], GPT-4, NewBing, medllama3-v20 [22]) to forecast interactions with potential or experimental drugs. From this, we generated 5,092 valid drug predictions, which were then ranked and interpreted. Additionally, AutoDock Vina facilitated 15,735 molecular docking experiments to corroborate LLM predictions. By conducting a quantitative analysis of LLM outputs and molecular docking outcomes, we assessed the efficacy and accuracy of LLMs in drug repositioning efforts. Notably, through experimental validation,

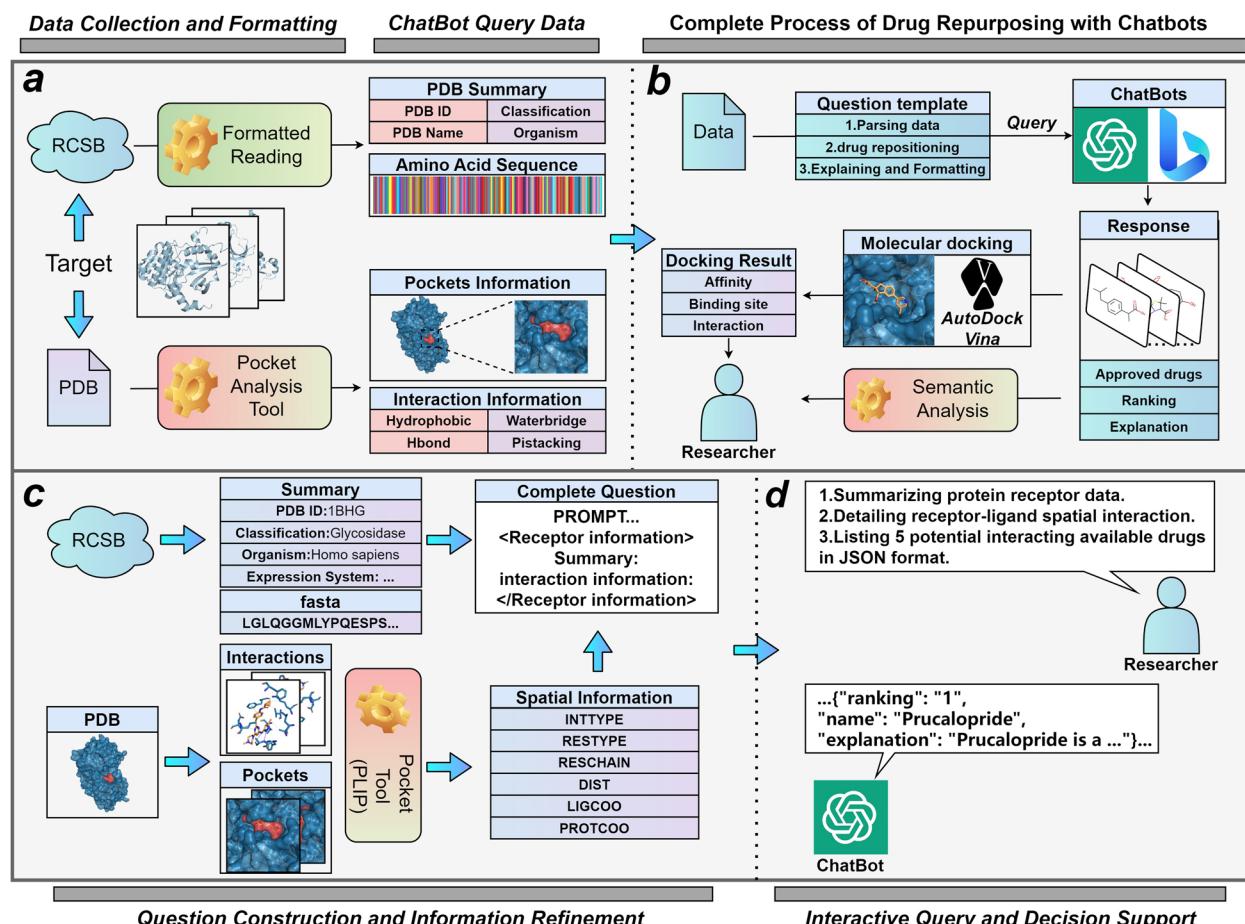
we identified two previously unrecognized drug-target interactions, promising significant advancements in cancer therapy. The principal contributions of our research are as follows:

1. We propose a drug repositioning framework based on LLMs and the integration of multisource prompts, aimed at efficiently exploring the full potential of existing drugs.
2. By incorporating spatial interaction information of small molecule ligands into the prompts, we significantly enhance the performance of LLMs in drug repositioning tasks.
3. Our analysis uncovers a strong correlation between LLMs' interpretation of target data and the outcomes of drug repositioning, indicating the potential for further enhancing LLM performance.

## Results

### DrugReAlign framework

To optimize the utilization of existing drugs, our research introduces an innovative drug repositioning framework that harnesses LLMs and incorporates multi-source prompting techniques, illustrated in Fig. 1. This framework is designed to enable LLMs to perform comprehensive analyses of target sites, facilitating self-prompting—a strategy proven effective in prior studies [23, 24]. One critical challenge with LLMs is the risk of hallucinations and misinformation, particularly when prompts are derived from a single or narrow data source. This can lead to inaccurate predictions or fabricated relationships [25, 26]. To address this, we employ a multi-source prompting approach that integrates diverse and reliable data inputs. In Additional File 1: Figure S1 [26, 27], we provide the prompt template and a template for the predicted results of drug repositioning executed



**Fig. 1** DrugReAlign Framework and Detailed Flowchart. **a** Construction of multisource prompts for targets, **(b)** Screening of potential drugs interacting with targets based on LLMs and multisource prompt information. **c** Example of Prompt Construction for Specific Targets **(d)** Interactive querying and decision support using LLMs

using the proposed framework. Additionally, to validate the usability of DrugReAlign on large-scale datasets, we tested its multithreading capabilities. The results, demonstrating the rapid screening ability of multithreaded DrugReAlign for large volumes of data, can be found in Additional File 1: Table S1.

Our methodology combines information from Protein Data Bank (PDB) structure summaries and known spatial interactions between targets and small molecules. By incorporating structural and interaction data from multiple sources, we significantly reduce the likelihood of LLMs generating misleading or false predictions. Specifically, the detailed structural data from PDB provides a factual and constrained foundation, while the interaction information from known small molecule relationships further narrows the model's focus to relevant biological interactions. This ensures that LLMs are guided by concrete and experimentally validated data, rather than relying solely on their internal language-based knowledge, which can sometimes lead to erroneous or non-existent predictions.

In module (a), we retrieved the PDB structure descriptions of targets from the RCSB [28] database, including PDB name, classification, gene ownership, etc. Additionally, we extract known spatial interaction information between targets and small molecules (such as spatial coordinates and interaction forces) from the PDB files to form example-based prompts. This multi-source prompt construction provides the LLMs with real-world, experimentally verified examples, ensuring that its outputs are grounded in factual data. Module (c) demonstrates the construction of example-based prompts for specific targets, and rationales. In module (b), the summary of PDB structures and target-small molecule spatial interactions are inputted into LLMs. This combined data ensures that the LLM is working within a well-defined, reliable context, which helps to mitigate the risk of hallucinations. The LLM analyzes these prompts and generates suggestions for drug repositioning, focusing on interactions that are biologically plausible and supported by data.

Subsequently, we extract the predicted drug structures and conduct molecular docking experiments with the

respective targets to validate the proposed framework's effectiveness in drug repositioning tasks. In module (d), using LLMs and multi-source prompt techniques, the framework generates detailed analysis reports, including rationale for drug rankings, molecular docking results, and interactive queries, enhancing the transparency and reliability of the process. This combination of multi-source prompts and experimental validation helps further mitigate hallucinations by providing data-driven outputs at every stage, ensuring that predictions remain within the bounds of known science. Thus, this framework not only provides target analysis based on LLMs but also minimizes the risks associated with hallucinations and misinformation through its reliance on multi-source prompting, improving the reliability of potential drug repositioning relationships. This approach effectively alleviates the researcher and resource-intensive challenges of traditional methods in drug repositioning tasks.

#### Validation of DrugReAlign on DTI datasets

To preliminarily validate the reliability of DrugReAlign and LLMs in the task of drug repositioning, we selected NR [29] and GPCR [29] as DTI benchmark datasets for validation. Detailed information about these datasets can be found in Additional File 1: Table S2. The results of individual runs from the five repeated experiments are provided in Additional File 2. Initially, we identified the PDB structures corresponding to the relevant targets and then used DrugReAlign to search for market-available drugs that might interact with these targets. The specific experimental results are presented in Table 1.

**Target Coverage Rate (TCR):** The proportion of targets with at least one drug interaction recorded in the dataset.  
**Top1 Recommendation Success Rate (T1RSR):** The proportion of targets where the top1 recommended drug is recorded in the dataset.  
**Overall Interaction Rate (OIR):** The proportion of all drug recommendations that are recorded in the dataset. Metrics are averaged over 5 runs, and Bold values represent the highest performance, with standard deviations included to reflect variability. NR contains 25 targets and GPCR contains 68 targets.

**Table 1** Performance of DrugReAlign in drug recommendations within the DTI datasets

Datasets	Metric	GPT-4	New Bing	GPT-3.5	medllama3-v20
NR [29]	TCR↑	<b>0.776(0.054)</b>	0.736(0.020)	0.496(0.041)	0.360(0.025)
	T1RSR↑	<b>0.328(0.064)</b>	0.296(0.086)	0.312(0.082)	0.208(0.053)
	OIR -	0.232(0.021)	0.224(0.032)	0.176(0.020)	0.104(0.024)
GPCR [29]	TCR↑	<b>0.641(0.061)</b>	0.514(0.025)	0.618(0.021)	0.418(0.020)
	T1RSR↑	<b>0.279(0.073)</b>	0.244(0.080)	0.262(0.077)	0.235(0.074)
	OIR -	0.176(0.010)	0.183(0.012)	0.181(0.009)	0.131(0.011)

We first defined the TCR metric, which reflects the proportion of recommended results that include known DTI relationships. This metric is significant for the drug repositioning task because DrugReAlign requires LLMs to analyze target information to complete the recommendation task. The presence of known DTIs indicates that LLMs have, to some extent, understood the conditions for interacting with the target, thereby greatly enhancing the credibility of the recommendation results. Additionally, we defined the T1SR metric, which represents the proportion of recommendations where the top result is a known DTI among all recommended results. This metric similarly indicates the credibility of the recommendation results. OIR represents the proportion of known DTI relationships among all recommended results. An increase in this metric indicates an improvement in the credibility of DTI recommendations, while a decrease might indicate an increase in the diversity of recommendation results. Therefore, in different scenarios, we need to consider both the credibility and diversity of the recommendation results to achieve a balance between the two.

In both datasets, GPT-4 achieved the best performance in the TCR and T1SR metrics, while GPT-3.5 exhibited significant fluctuations in both datasets. This may be due to the varying proportions of different target types in GPT-3.5's training data. Additionally, NewBing showed relatively lower TCR and T1SR values in both datasets, indicating a tendency to recommend unknown DTIs. However, NewBing's OIR in the GPCR dataset was comparable to other models, suggesting that while it focused on known DTIs for certain targets, it explored unknown DTIs for others, offering a balance between discovery and reliability.

As for medllama3-v20, although its performance was lower in both TCR and T1SR metrics, the decline in OIR might suggest a greater focus on exploring novel DTI relationships, thereby enhancing diversity in its recommendations. This could provide potential for discovering new interactions, despite a lower immediate credibility in known DTI coverage. Overall, all models performed relatively well in terms of TCR, with most recommendations including known DTIs. This strongly demonstrates the robustness of LLMs in the task of drug repositioning, with varying degrees of focus on known versus novel DTI discovery.

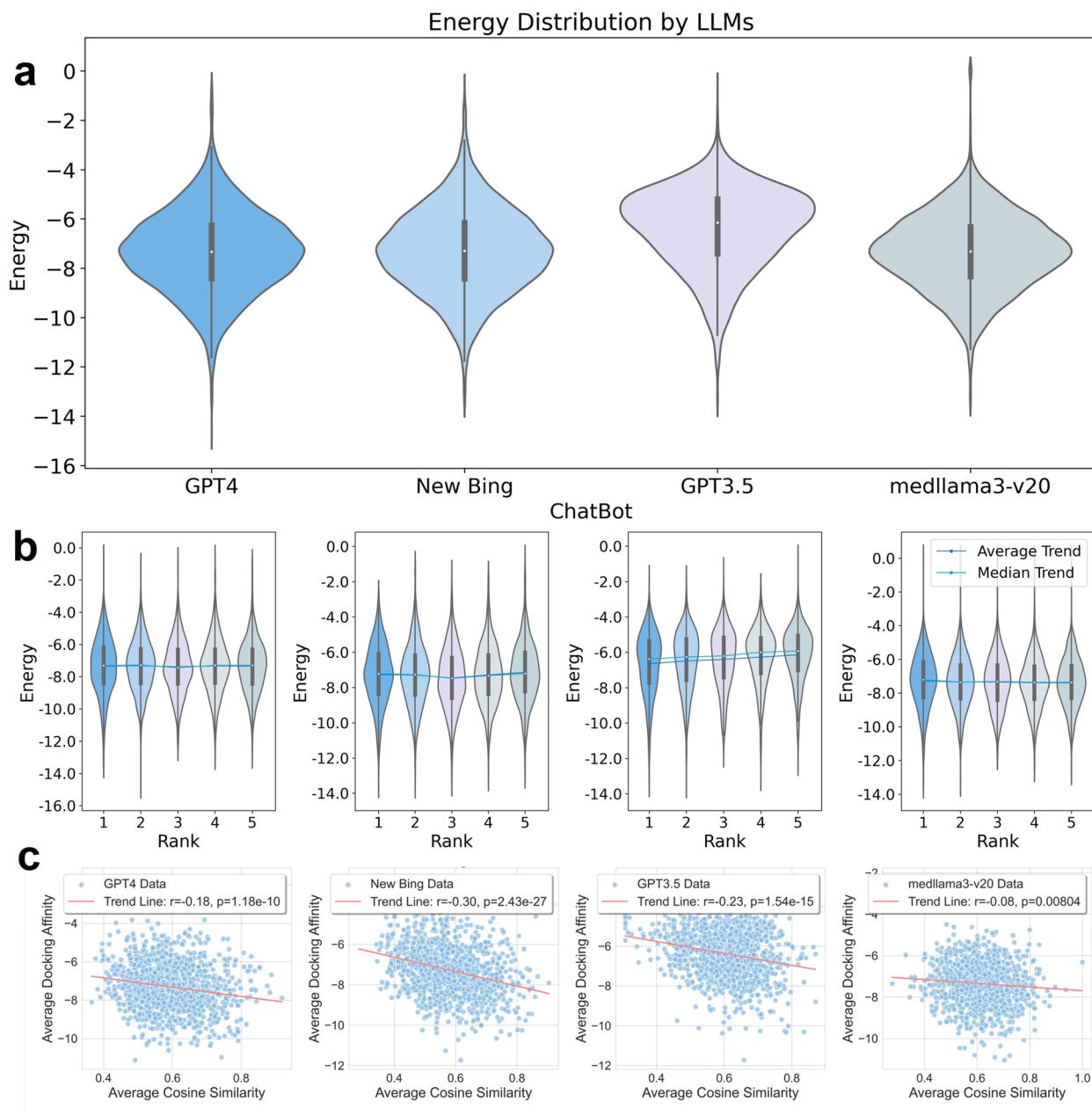
#### Quantitative analysis of drug repositioning results based on AutoDock Vina

Molecular docking technology is extensively utilized to evaluate and screen candidate drugs for targets. Molecular docking software can predict their binding affinity by simulating the interaction between a drug and its target.

The value of the affinity largely indicates a drug's binding capability and potential biological activity towards a specific target. The higher the affinity, the tighter the drug binds to the target, potentially indicating stronger biological effects and better therapeutic potential [30]. To assess the prediction results of LLMs, we have conducted a substantial number of molecular docking experiments using AutoDock Vina on relevant targets and drugs. In molecular docking, binding free energy lower than -5 kcal/mol generally indicates strong molecular interactions, while values lower than -7 kcal/mol or beyond are regarded as strong interactions.

Figure 2(a) and Table 2 display the molecular docking scores for drugs and their corresponding targets as predicted by GPT-3.5, GPT-4.0, New Bing model and medllama3-v20 models. We observed that the results predicted by all three LLMs models achieved satisfactory molecular docking scores, indicating their capability to predict potential therapeutic drugs for specific targets. The predictions by GPT-4 and New Bing models scored higher in docking experiments, whereas GPT-3.5's predictions were significantly weaker, with an average binding free energy of -6.40 kcal/mol. Interestingly, despite having only 8B parameters, medllama3-v20's average docking score (-7.35 kcal/mol) was comparable to that of GPT-4, showing promising performance. This suggests medllama3-v20's potential to deliver competitive results in molecular docking, though its consistency remains an area for further exploration.

These findings suggest a certain correlation between molecular docking scores and the overall capability of LLMs, potentially offering a quantitative method to assess the performance of LLMs in drug repositioning tasks. Further, we explored the correlation between the drug rankings predicted by LLMs and their molecular docking scores, as shown in Fig. 2(b) for GPT-4 and New Bing models, the average trend lines were relatively flat, indicating no significant linear relationship between the rankings of predicted drugs and the distribution of docking scores. Notably, the GPT-3.5 model was significantly behind other LLMs in overall performance, yet the rankings of predicted drugs showed an approximate linear relationship with the distribution of docking scores. For medllama3-v20, the performance in drug rankings and docking scores was quite similar to that of GPT-4, which might be surprising given its 8B parameter size. This suggests that medllama3-v20, despite having fewer parameters, can still achieve results comparable to larger models like GPT-4 by focusing on similar aspects when parsing instructions. This highlights that different LLMs may prioritize different elements of the task, with medllama3-v20 demonstrating capabilities on par with



**Fig. 2** Distribution of binding free energy and semantic relevance analysis for drugs recommended by LLMs. **a** Distribution of binding free energy between drugs predicted by LLMs and their corresponding targets. **b** Correlation between the rankings of drugs predicted by LLMs and binding free energy. **c** Exploration of the correlation between average binding free energy and semantic similarity for drugs targeting the same site

more advanced models in abstract tasks, while other models may perform better with more direct and concrete instructions.

#### Correlation between interpretability and docking performance of LLMs

LLMs can provide precise answers on specific topics through concise dialogue. In this study, LLMs output predicted drug names, rankings, and reasons based

on multi-source cue information of the target. In this section, we particularly focus on the reasons (or explanations) provided by LLMs for predicting drugs and evaluate the quality of explanations given by LLMs when predicting drugs for specific targets. We used OpenAI's text embedding model, text-embedding-3-small, to convert all predicted drug explanations into vectors for quantitative assessment.

**Table 2** Comparison of binding free energy (kcal/mol) for drugs recommended by LLMs based on target information. Each model was tested using 1,278 targets to evaluate their performance

Models	Average score	Median score	Minimum score
GPT4	<b>-7.35</b>	<b>-7.33</b>	<b>-14.74</b>
New Bing	-7.29	-7.29	-13.43
GPT3.5	-6.40	-6.14	-13.43
medllama3-v20	<b>-7.35</b>	-7.32	-13.42

Specifically, we quantified the correlation between the explanations for predicted drugs by LLMs and the molecular docking scores. We selected the drug information for each target repositioning as a unit, calculated the average cosine similarity of the explanations for predicted drugs within these units, and calculated the average binding free energy of the predicted drugs to their respective targets. Based on all units, we calculated the correlation coefficient and p-value between the average cosine similarity of drug explanations and the corresponding average molecular docking scores to quantify the strength and significance of their relationship. As shown in Fig. 2(d), for the three types of LLMs, the calculated correlation coefficients were all less than 0. In the New Bing model, the correlation coefficient reached -0.3, indicating a slight to moderate negative correlation trend. That is, there is a positive correlation between the similarity of explanations for predicted drugs and their molecular docking affinity. In other words, the higher the similarity of LLM explanations for drugs predicted for a specific target, the stronger the binding affinity of these predicted drugs with that target. Moreover, in all LLMs, the calculated p-values were significantly less than 0.05, indicating that the negative correlation between the average cosine similarity of drug explanations and the corresponding average molecular docking scores is statistically significant. For medllama3-v20, the correlation was noticeably weaker compared to other LLMs. After reviewing its explanations, we found that medllama3-v20 provided simpler and more uniform responses for different drug predictions. This lack of detail may be due to its smaller parameter size (8B), limiting its ability to capture nuanced relationships between drugs and targets. Additionally, it might have been fine-tuned on instruction-based datasets, which could focus more on following general patterns rather than providing diverse and detailed explanations. As a result, both its simplicity and potential focus on instructions likely contributed to the weaker correlation observed.

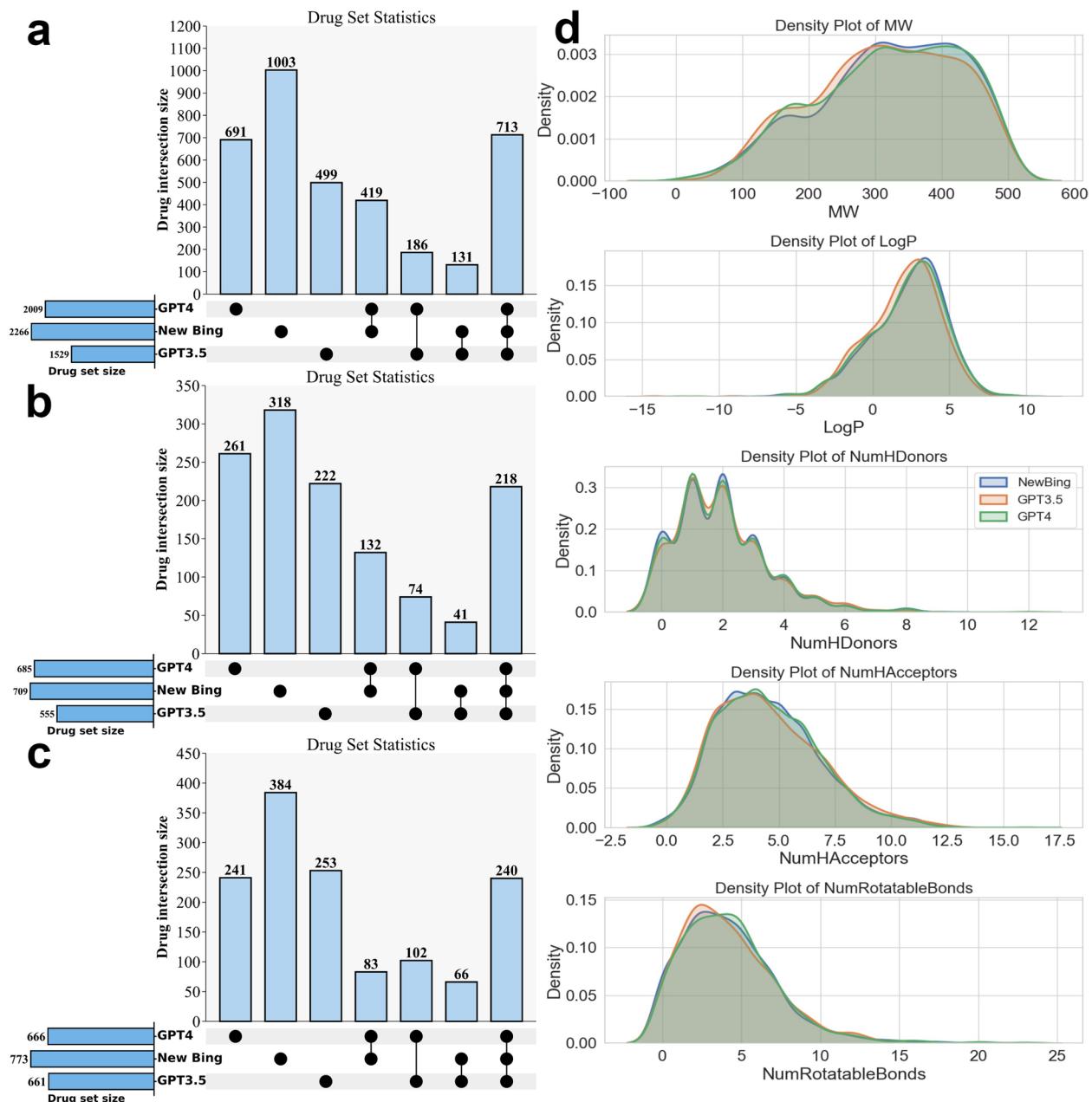
#### Propensity of predicted drugs

In this section, we evaluated the tendencies of the three types of LLMs in drug repositioning tasks within the

proposed framework. Figure 3(a) shows the intersection size of the drug sets predicted by the three models. The New Bing model predicted a total of 2,266 drugs, surpassing other LLMs and indicating its broader coverage in the drug database. It is also observed that 1,003 drugs were exclusively predicted by the New Bing model, suggesting its ability to provide a greater number of alternative drugs not covered by other models. The GPT-3.5 model predicted only 1,424 drugs, significantly fewer than the other LLMs, with just 499 drugs being unique predictions. We speculate that this is due to differences in the volume of training data sources among the LLMs. Therefore, when utilizing LLMs for drug repositioning tasks, it is crucial to consider the breadth and update frequency of their training data. This can offer more potential drugs for specific research targets, helping to expand existing treatment options and develop new ones.

We utilized the Lipinski's Rule of Five [31] (molecular weight (MW), lipophilicity (LogP), number of hydrogen bond donors (HBD), number of hydrogen bond acceptors (HBA), and number of rotatable bonds (RotB)) to analyze and evaluate predicted small molecule drugs with a molecular weight less than 500. As observed in Fig. 3(d), compared to the GPT-4 and New Bing models, the GPT-3.5 model exhibited a clear preference for drugs with smaller molecular weights. Furthermore, we also analyzed and evaluated large molecule drugs predicted by the proposed framework with molecular weights greater than 500. In the study, the New Bing and GPT-4 models predicted 385 and 354 large molecule drugs, respectively, while GPT-3.5 predicted only 256. This further emphasizes the GPT-3.5 model's preference for small molecule drugs.

The preference of the GPT-3.5 model also extends to the chemical properties of the selected drugs. In Fig. 3(d), the GPT-3.5 model more frequently outputs small molecule drugs with relatively lower LogP and RotB values, and the number of HBDs is usually slightly higher. A higher number of HBDs enhances the molecule's hydrophilicity, at this time with lower lipophilicity and higher rigidity, consistent with experimental outcomes. These findings reflect the GPT-3.5 model's tendency to select drugs with smaller molecular weights, stronger hydrophilicity, lower lipophilicity, and higher rigidity. Enhanced hydrophilicity aids in the solubility and absorption of the drug, while the rigidity of the molecular structure helps in precise binding with the target, which are key factors to consider in drug development. In Additional File 1: Figure S2 [32], we provide the ADMET property analysis of LLMs-recommended drugs executed using the proposed framework, which enables a more comprehensive analysis of the differences among these drugs.



**Fig. 3** Relationships among recommended drug sets and distribution of Lipinski's rule properties of LLMs. **a** Relationships among drug recommendation sets of different LLMs under the complete dataset. **b** Relationships among drug recommendation sets of different LLMs under subsets of data. **c** Relationships among drug recommendation sets of different LLMs under subsets of data with spatial interaction information removed. **d** Distribution of Lipinski's rule properties of drugs recommended by different LLMs under the complete datasets. The bars represent the number of drugs in each intersection, with their height indicating the intersection size. The dots below show which drug sets are being intersected, and connected dots represent drugs shared across multiple sets. The bars on the left display the total drug set size for each model

#### Ablation experiment

We conducted ablation studies to investigate whether spatial interaction information between drugs and targets can serve as an effective prompt to enhance the drug

repurposing performance of LLMs, as shown in Table 3. Here, "w/o SF" indicates the exclusion of spatial interaction data between drugs and targets as prompt information and the removal of corresponding analysis requests. We randomly selected 218 targets and reapplied three

**Table 3** Comparison of binding free energy (kcal/mol) between LLMs before and after the removal of drug-target spatial interaction information and traditional deep learning models

Models	Average score	Median score	Minimum score
GPT3.5	-6.41	-6.19	<b>-13.43</b>
GPT3.5 w/o SF	<b>-6.88</b>	<b>-6.75</b>	-11.91
GPT4	<b>-7.36</b>	<b>-7.29</b>	<b>-14.74</b>
GPT4 w/o SF	-6.85	-6.73	-12.79
New Bing	<b>-7.29</b>	<b>-7.29</b>	<b>-13.43</b>
New Bing w/o SF	-6.93	-6.78	-12.79
medllama3-v20	<b>-7.43</b>	<b>-7.43</b>	<b>-13.47</b>
medllama3-v20 w/o SF	-7.29	-7.40	-13.42
DrugBan	-6.02	<b>-6.10</b>	<b>-10.83</b>
TransformerCPI2.0	<b>-6.07</b>	-6.04	-9.52

types of LLMs for drug repurposing, using AutoDock Vina to calculate the binding free energy.

An intriguing observation is that both GPT-4 and New Bing models show a notable decline in molecular docking scores when information on drug-target spatial interactions is removed. In contrast, the docking scores for the GPT-3.5 model significantly improve. With the removal of drug-target spatial interaction cues, the molecular docking scores of all three models align at the same baseline level. These results clearly indicate that GPT-3.5's ability to comprehend drug-target spatial interaction information is limited. Information not understood and perceived as noise has a noticeable negative impact on the outcomes of drug repositioning [33]. However, for simple target summary information, the comprehension abilities of the models are largely similar, showing no significant differences. These findings highlight the importance of tailoring prompts to leverage the strengths of different LLMs for achieving optimal prediction outcomes. For medllama3-v20, the removal of spatial interaction information led to a relatively smaller decline in docking scores compared to GPT-4 and New Bing. Its performance without spatial cues (-7.29 kcal/mol) remained close to its original score, indicating that medllama3-v20 might rely less on spatial interaction data than other models. This could be attributed to its design or fine-tuning approach, potentially focusing more on general structural or sequence-based information. As a result, medllama3-v20 maintained more stable predictions, suggesting it may be more resilient in scenarios where spatial interaction data is unavailable or incomplete.

Figure 3(b) displays the distribution of predicted drug quantities by the three LLMs when retaining drug-target spatial interaction information. Figure 3(c) shows the

distribution of predicted drug quantities by the three models after removing this information. When spatial cues are omitted, both the total quantity of drugs predicted by GPT-3.5 and the number of unique predictions significantly increase, further supporting that drug-target spatial interaction information may be perceived as noise within the GPT-3.5 model. In contrast, the total number of drugs predicted by the GPT-4 model remains relatively unchanged, but there is a noticeable decrease in the number of unique drug predictions, suggesting that the novelty of drug predictions by the GPT-4 model could be compromised without spatial cues. For the New Bing model, both the total and unique numbers of predicted drugs significantly increase after spatial cues are removed. However, according to Table 3, the average molecular docking scores for drugs predicted by the New Bing model and their corresponding targets significantly decrease. This may be because drug-target spatial interaction information acts as an effective constraint, enhancing the predictive performance of the New Bing model. With the removal of this spatial interaction information, constraints are relaxed, potentially leading the New Bing model to predict a greater number of less effective drugs.

Furthermore, to compare the performance of traditional deep learning models with LLMs in drug repurposing tasks, we selected two state-of-the-art (SOTA) models, DrugBan [34] and TransformerCPI2.0 [35], to conduct experiments on the previously selected 218 targets. Specifically, we used 14,642 compounds from the BindingDB dataset organized by Bai et al. as experimental content. We predicted the interactions between the targets and all drugs, selecting the top 5 drugs with the highest scores as the drug repurposing results for each target, followed by molecular docking and corresponding analysis, with the specific binding free energy results as shown in Table 3.

The experimental data indicate that traditional deep learning models significantly underperform LLMs in drug repurposing tasks, with the two SOTA models achieving an average binding free energy of only about -6 kcal/mol. Additionally, deep learning models also significantly lack novelty in drug repurposing; among the 14,642 candidate compounds, only 285 compounds are present in the DrugBan model's predictions, and 561 in TransformerCPI2.0's predictions, while a single LLM's recommendations can reach more than 700 compounds. We believe this is primarily due to the significant difference between the target domain of drug repurposing and the source domain of the training data, with relevant targets not being present in the model's training data, thus leading to poor model performance. The cross-domain issue is a common shortcoming of single-task models designed

for specific tasks, mainly limited by the training data and the model's parameters. In contrast, LLMs effectively overcome these deficiencies. The vast amount of training data and model parameters enable LLMs to exhibit good generalization capabilities across various tasks.

### Evaluating LLM-Driven drug repositioning with deep learning models

Deep learning models are widely used in fields such as drug repurposing and drug-target interactions, but they are constrained by the available training data and generally lack strong generalization capabilities for unseen data. However, with LLMs extensive training data and large number of parameters, hold promise for overcoming these limitations. To clarify the constraints of deep learning,

models on related issues, we utilized Transformer-CPI2.0 and DrugBan to predict DTIs based on drug repurposing results predicted by LLMs, as shown in Fig. 4. Most data points in the graphs deviate from the diagonal line, indicating significant differences in the predictions of the two models, and most data points were deemed non-interacting, which contradicts the results of large-scale molecular docking studies. These experimental results suggest that the drug recommendations predicted by LLMs are novel and difficult to predict with traditional deep learning methods. Thus, LLMs can indeed significantly break through the limitations of traditional models in terms of training data and parameters, and they hold great significance for the field of drug discovery.

Besides, to further explore the reliability of LLMs in drug repurposing recommendations, we filtered experimental protein targets based on a sequence similarity greater than 90% (PDB IDs: 1EQZ: Entity 4 [36], 3AFA: Entity 1 [37], 1U35: Entity 2 [38], 3LEL: Entity 1 [39]) for our experiments. We collected drug repurposing results for these four protein targets from different LLMs (GPT4, GPT3.5, NewBing). Then, using MolFormer [27], we embedded the drugs obtained and performed dimensionality reduction with the T-SNE algorithm.

As shown in Fig. 4(d), the majority of the drugs recommended by the LLMs cluster in a small feature space, indicating that the drugs recommended by LLMs for protein targets with similar sequences exhibit high similarity in their feature space. This suggests that these drugs likely share similar pharmacological effects and chemical structures. This experimental result also supports the credibility of drug recommendations made by LLMs.

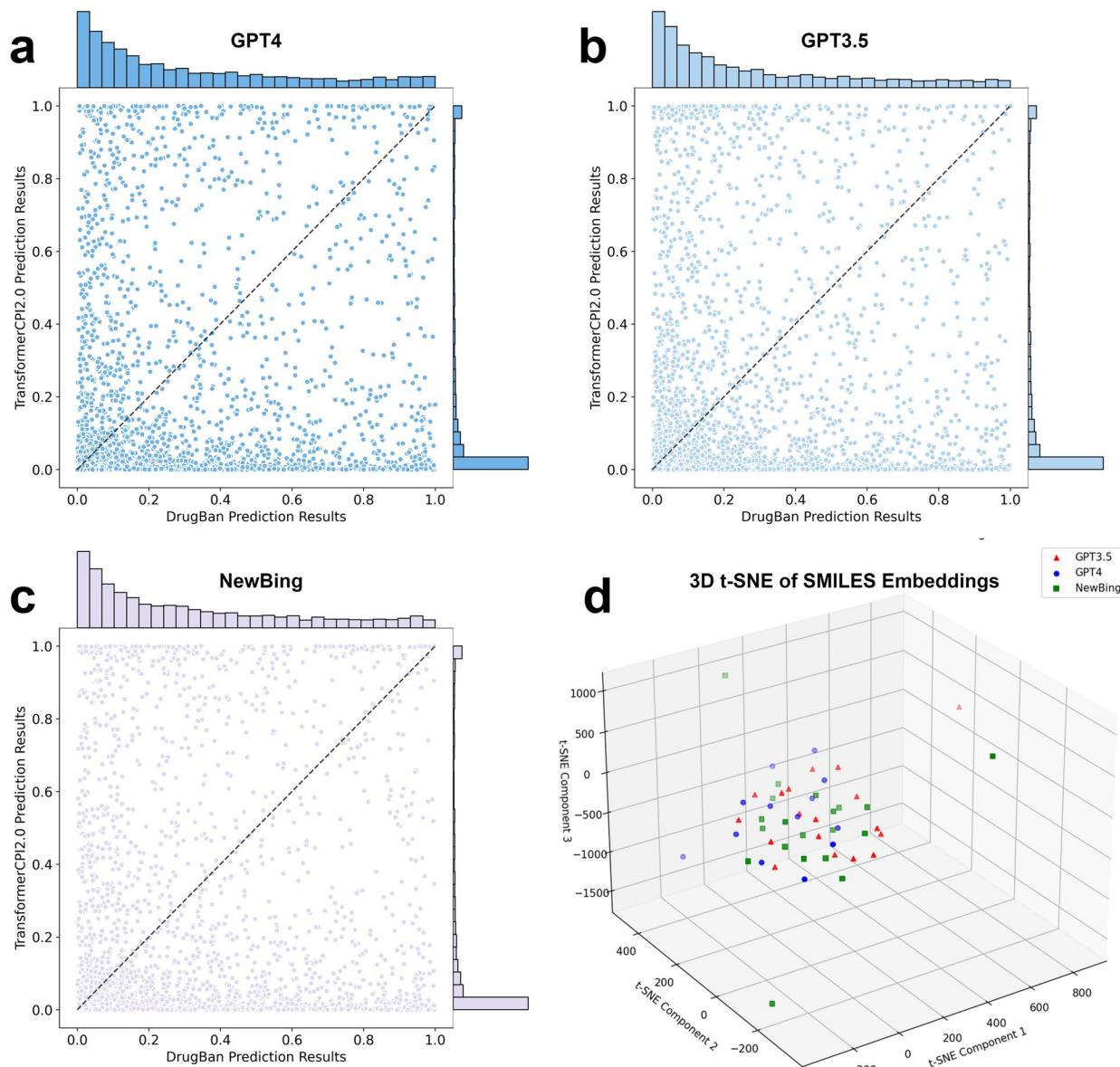
### Case study

To further validate the reliability of DrugReAlign in drug repositioning tasks, we selected PARP1 (Poly (ADP-ribose) polymerase 1) for case analysis. PARP1, a key

DNA repair enzyme, plays a critical role in cancer treatment, particularly by inhibiting its activity to prevent cancer cells from repairing DNA damage, leading to cell death. This approach is especially effective for cancers with DNA repair defects, such as BRCA-mutated breast and ovarian cancers. In addition, PARP1 inhibitors can enhance the effects of chemotherapy and radiotherapy, promote immune response, and even show therapeutic potential in cardiovascular and neurodegenerative diseases. As a target, PARP1 demonstrates broad therapeutic applications across multiple diseases. However, current PARP inhibitors face challenges such as resistance and lack of selectivity, highlighting the need for new drugs to more effectively regulate PARP1 interactions, improve therapeutic outcomes, overcome resistance, and explore its potential in treating other diseases.

We first used the GPT-4-based DrugReAlign framework to obtain drug repositioning results and selected the top two recommended drugs (Midostaurin and Dasatinib) for molecular docking. Subsequently, molecular dynamics (MD) simulations were performed using GROMACS [40] until the system stabilized. The specific parameter settings can be found in the supplementary materials. As shown in Fig. 5(a), the left panel displays the initial docking results, while the right panel shows the final outcomes after the MD simulations. We found that after the MD simulations, both drugs formed tighter interactions with PARP1, penetrating deeper into the binding pocket.

Additionally, Fig. 5(b) illustrates the RMSD changes of the drug-target systems during the MD simulations. The RMSD curve reflects the deviation of the drug molecules from their initial docking conformation, helping assess the stability of drug-target binding and structural changes. From the figure, we observe that Midostaurin was very stable during the initial phase of the simulation, underwent significant conformational adjustments after 20 ns, and around 35 ns, the RMSD values decreased again, indicating that its structure approached its initial state. The system then entered a long-term stable phase, which may reflect exploratory movements of the drug within the binding pocket before settling into a stable binding mode. This suggests that Midostaurin formed stable interactions with PARP1, indicating strong binding affinity and relatively low dissociation rates, with stability maintained even in a dynamic environment. Moreover, as shown in Fig. 5(c), the number and occupancy of hydrogen bonds in the system reached their peak after 35 ns, and the number of  $\pi$ - $\pi$  stacking and hydrophobic interactions shown in Fig. 5(d) also increased after 35 ns. These strong non-covalent interactions indicate that Midostaurin's binding to PARP1 is quite stable. In contrast, Fig. 5(b) shows that Dasatinib underwent significant conformational changes



**Fig. 4** Analysis of deep learning predictions for LLM-recommended drugs and dimensionality reduction visualization of pre-trained molecular representation models. **a-c** Drug-target interaction predictions for drugs recommended by GPT-4, GPT-3.5, and NewBing, respectively. **d** Visualization of dimensionality reduction using T-SNE on drug recommendations from LLMs

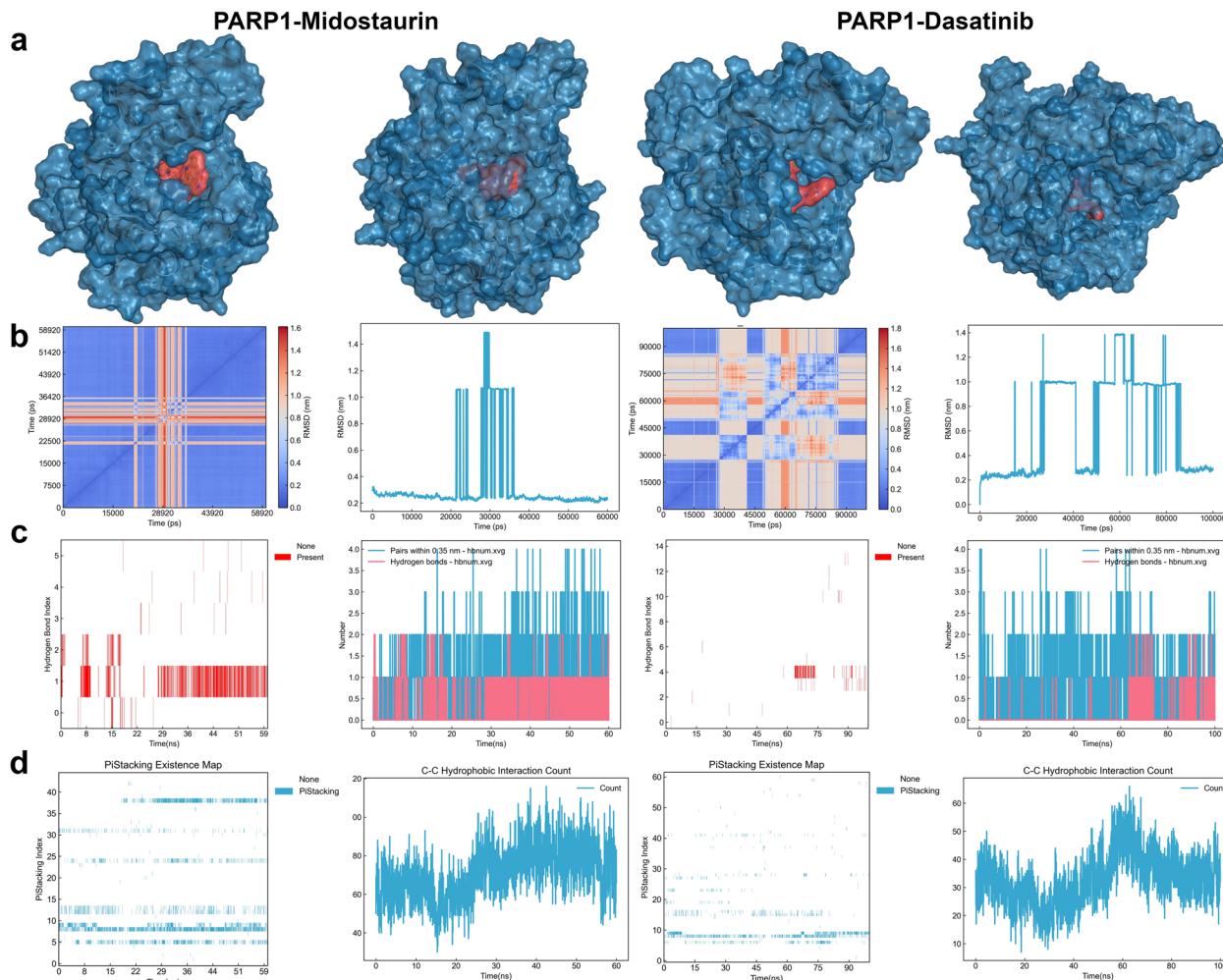
during the simulation, reaching near steady-state around 85 ns, indicating that its initial docking conformation was not the optimal binding site. As shown in Fig. 5(c), the number of hydrogen bonds between the drug and target began to increase around 65 ns, but at this point, the RMSD values were still high, suggesting that the drug had not yet reached a stable state and was still exploring conformations within the binding pocket. Around 85 ns, the number of hydrogen bonds peaked again, forming a relatively stable binding state. The  $\pi$ - $\pi$  stacking and hydrophobic interactions in Fig. 5(d) also stabilized around

this time. In summary, both Midostaurin and Dasatinib demonstrate strong potential interactions with the target PARP1, with Midostaurin showing relatively high stability throughout the simulation. These results strongly support the effectiveness of DrugReAlign in identifying potential drug repositioning candidates.

## Methods

### Data preparation

To assess the performance of the proposed framework in the area of targeted drug repositioning, we meticulously



**Fig. 5** LLM-Based Drug Repositioning and Molecular Dynamics Analysis of PARP1 Interactions with Midostaurin and Dasatinib. **a** Left: Initial binding site from molecular docking; Right: Final binding site from molecular dynamics simulation. **(b)** Left: RMSD matrix of the target-drug complex; Right: RMSD variation plot of the target-drug complex over time. **(c)** Left: Hydrogen bond occupancy of the target-drug complex over time; Right: Number of hydrogen bonds in the target-drug complex over time. **(d)** Left: Occupancy of  $\pi$ - $\pi$  interactions in the target-drug complex over time; Right: Number of hydrophobic interactions in the target-drug complex over time.

selected a subset of targets from the public database BindingDB to construct our dataset. The selection criteria required that each target have records of at least two small molecule ligands in the RCSB database [28] to ensure sufficient information on target-small molecule drug binding pockets. After data cleaning, we collected the PDB structures of 1,273 targets and their corresponding small molecule drugs, serving as the dataset for molecular docking. In this process, we also gathered detailed descriptive data about the targets, including PDB ID, protein sequence (in fasta format), PDB name, classification, biological source, expression system, etc., and extracted summary information. Additionally, we used the PLIP (Protein–Ligand Interaction Profiler) [41]

tool to analyze the interactions between targets and small molecule ligands and extracted information on target-small molecule drug spatial interactions. These two types of information were used as prompts for inputting into LLMs.

#### Target-ligand spatial interaction information

Analyzing the three-dimensional structural data of target-small molecule ligand interactions can facilitate a deeper understanding of the mechanisms of molecular interactions, guiding drug design and optimization. In this study, each target has at least two small molecule ligands with verified interactions. Specifically, we employed the PLIP (Protein–Ligand Interaction Profiler) [41] tool to analyze

protein-small molecule ligand interactions and extracted corresponding spatial interaction information such as atomic coordinates and bonds. This tool can accurately identify and classify various interaction types including hydrogen bonds, hydrophobic interactions, and  $\pi$ - $\pi$  stacking, as shown in Table 4. The process of obtaining and analyzing the three-dimensional structural information of targets and small molecule ligands takes into account the spatial relationships and chemical properties between atoms, ensuring the accuracy of spatial interaction identification. This ensures sufficient target-small molecule drug spatial interaction data, providing rich and precise constraining prompts for LLMs within the proposed framework, thereby enhancing drug repositioning performance.

### Molecule docking

Molecular docking simulates the interaction between proteins and small molecules or other biomacromolecules to predict the optimal binding mode of small molecules at the active site of the protein and the corresponding binding affinity. In drug repositioning tasks, we primarily employ AutoDock Vina software to perform molecular docking, thereby verifying whether there are interactions between the drugs predicted by the proposed framework and their respective targets.

Molecular docking primarily comprises two modules: a search algorithm and a scoring function. The search algorithm module is tasked with exploring all potential spatial positions and orientations of the drug molecule within the target protein's binding pocket. In this research, we utilize the method proposed by Saberi Fathi et al. [42] to locate all possible binding pockets, followed by an exhaustive search to identify the optimal binding pose. During this process, the flexibility of molecules, i.e., potential conformational changes of drug molecules and the protein during their interaction, needs to be considered. The search algorithm module aims to efficiently find the lowest energy binding mode within the vast conformational space.

The scoring function of AutoDock Vina software is used to estimate the binding affinity for each possible pose, assessing the stability of the drug molecule's binding to the target. The scoring function integrates various interaction forces such as hydrophobic interactions, hydrogen bonds, Van der Waals forces, and electrostatic interactions to calculate the energy value

of the drug molecule binding to the target. A lower value indicates tighter binding and stronger interaction between the drug and the target, suggesting higher potential biological activity of the drug. The Vina scoring function used in this study is:

$$V_{\text{score}} = E_{\text{vdW}} + E_{\text{HBond}} + E_{\text{electrostatic}} + E_{\text{desolvation}} \\ + E_{\text{hydrophobic}} + E_{\text{torsional}} + E_{\text{conformational}}, \quad (1)$$

where  $E_{\text{vdW}}$  represents Van der Waals energy, which describes the attraction or repulsion between molecules based on their proximity.  $E_{\text{HBond}}$  denotes hydrogen bond energy, quantifying the contributions of hydrogen bonds in molecular interactions, vital for specificity among biomolecules.  $E_{\text{electrostatic}}$  represents electrostatic energy, reflecting the forces of attraction or repulsion among charged molecules.  $E_{\text{desolvation}}$  denotes desolvation energy, indicative of the energy change when a molecule moves from a solvent environment to a binding site.  $E_{\text{hydrophobic}}$  represents hydrophobic interaction energy, which reflects the attraction between molecules arising from the hydrophobic effect.  $E_{\text{torsional}}$  indicates torsional energy, associated with the freedom of rotatable bonds in a ligand and potential conformational changes upon binding.  $E_{\text{conformational}}$  signifies conformation-dependent internal energy, evaluating the stability of interactions among various molecular segments.

### Semantic analysis

In this study, when commanding LLMs to output predicted drugs, they are also required to provide corresponding explanations. To quantify the semantic information in the explanations and explore its contribution to the prediction results, we utilize OpenAI's text Embedding model—text-embedding-3-small to transform the explanatory text  $T$  into a vector representation  $V$  in a high-dimensional vector space. This process primarily includes the following steps:

Tokenization: The input text  $T$  is first processed through tokenization, converting it into a series of tokens:

$$\{t_1, t_2, \dots, t_m\} = \text{Tokenize}(T) \quad (2)$$

**Table 4** Detailed information used for generating prompts in the PLIP tool

Types	Description	required attributes
hydrophobic	Repulsion between water and nonpolar substances	restype, reschain, restype_lig, dist, protcoo, ligcoo
hbond	Attraction between a hydrogen atom and electronegative atom	restype, reschain, restype_lig, protcoo, ligcoo, dist_h-a, dist_d-a, don_angle, donortype, acceptortype
pistackin	Interaction between aromatic rings	restype, reschain, restype_lig, centdist, angle, offset, protcoo, ligcoo

Embedding: Then, the text-embedding-3-small model  $E$  transforms these tokens into a 1536-dimensional vector, serving as the representation for the explanation corresponding to the predicted drug. This step can be expressed as:

$$V_{1536} = E(\{t_1, t_2, \dots, t_m\}) \quad (3)$$

Subsequently, we use cosine similarity to calculate the semantic similarity between explanations for predicted drugs:

$$\text{cosine\_similarity}(V_1, V_2) = \frac{V_1 \cdot V_2}{|V_1||V_2|} \quad (4)$$

where  $V_1$  and  $V_2$  are the text vectors corresponding to the explanations for drug molecules 1 and 2, respectively.  $V_1 \cdot V_2$  represents the dot product of  $V_1$  and  $V_2$ , and  $|V_1|$  and  $|V_2|$  respectively represent the Euclidean norms (i.e., the lengths) of vectors  $V_1$  and  $V_2$ . The cosine similarity within a single unit of drug repositioning results for each target can be represented as:

$$\text{unit\_cosine\_similarity} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{V_i \cdot V_j}{|V_i||V_j|} \quad (5)$$

where  $n$  is the number of drug explanations within the unit, and  $V_i$  and  $V_j$  are the embedding vectors corresponding to the explanations for the  $i$  and  $j$  drugs, respectively.

## Discussion

Drug repurposing offers significant benefits in saving time and resources, and this study highlights both the strengths and limitations of using deep learning and large language models (LLMs) for these tasks. While deep learning methods have demonstrated creativity, their reliance on specific datasets limits their generalizability, often leading to overfitting and less accurate real-world predictions.

LLMs like GPT-3.5, GPT-4, and New Bing, with their vast knowledge bases, overcome some of these limitations by offering a broader understanding of drug-target interactions. Our framework, DrugReAlign, enhances the predictive power of LLMs by integrating target summaries and spatial interaction information, resulting in improved accuracy and interpretability, as shown through molecular docking and validation on DTI datasets. However, challenges remain. LLMs can suffer from hallucinations, generating incorrect results when the input data is sparse or ambiguous. Our ablation studies highlighted the critical role of incorporating spatial interaction data to mitigate this issue and improve prediction robustness. Ensuring high-quality, multi-source

data is key to enhancing LLM-driven drug repurposing. In conclusion, while LLMs show great potential, further research is needed to refine their use in drug repurposing by improving prompt design, data quality, and handling the risks of hallucination.

## Conclusions

This research demonstrates the substantial value and potential of the proposed DrugReAlign framework in the field of drug repurposing. Through a series of extensive docking experiments with target molecules, we validated the efficient performance of prominent LLMs such as GPT-3.5, GPT-4, and New Bing in drug repurposing tasks. Our findings show a strong correlation between the semantic similarity of LLM-generated drug explanations and the analysis reports of their corresponding targets, further supported by the drug-target docking scores. This positive relationship underscores the importance of precise and comprehensive prompts when utilizing LLMs in these tasks.

Moreover, the framework not only enhances prediction accuracy but also deepens our understanding of drug action mechanisms and the complex interactions between drugs and their target proteins. By incorporating spatial interaction data and comprehensive target summaries, DrugReAlign offers a more refined and interpretative approach to drug repurposing. While the study demonstrates promising outcomes, it also sheds light on challenges faced by LLMs in practical applications, such as the need to further improve the reliability and interpretability of predictions. Future research could expand the application of LLMs into other biomedical fields and refine their prediction models to provide more reliable results. With continued advancements, these methods have the potential to revolutionize drug discovery, accelerating the process from laboratory findings to clinical trials, and ultimately leading to the development of safer and more effective treatments.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12915-024-02028-3>.

Additional file 1: Figures S1-S2 and Tables S1-S2. The document outlines key topics in drug repurposing with Large Language Models (LLMs), including questioning strategies, implementation details of DrugBan and TransformerCPI2.0, ADMET attributes of recommended drugs, computational resource usage, benchmark datasets NR and GPCR, molecular dynamics simulation parameters and molecular docking workflows. Figure S1: Drug repositioning question template. (a) Question prompt template, where the red part needs to be filled in with the target's information. (b) Target overview information. (c) Target sequence information. (d) Spatial interaction information between target and ligand. Figure S2: ADMET attributes of LLMs recommended drugs. Table S1: Runtime of DrugReAlign with different thread counts for processing 100 instances of PDB ID: 2EA2. Table S2: Drugs, targets, and interactions in each dataset.

Additional file 2: Repeated experiments. This file contains the specific data from five repeated experiments.

## Acknowledgements

We would like to thank the National Natural Science Foundation of China for their support of this work (No. 62450002, 62302339, 62372158). We also appreciate the contributions of all authors involved in this study, whose collaboration made this research possible.

## Authors' contributions

Jinhang Wei and Linlin Zhuo contributed to the initial draft and the design and implementation of the experiments; Xiangzheng Fu and Li Wang were responsible for data collection and reference preparation; XiangXiang Zeng, Quan Zou, and Dongsheng Cao provided experimental guidance and revised the manuscript. All authors read and approved the final manuscript.

## Funding

The work was supported by the National Natural Science Foundation of China (No. 62450002, 62302339, 62372158).

## Availability of data and materials

All data generated or analyzed during this study are included in this published article, its supplementary information files, and publicly available repositories. The sources of all analyzed datasets are as follows: The implementation of DrugReAlign is accessible via <https://github.com/kkkayle/DrugReAlign>, and the raw data can be downloaded from <https://github.com/kkkayle/DrugReAlign/tree/master/Experimental>. Additionally, the data is publicly available on Zenodo: <https://doi.org/10.5281/zenodo.13836105>.

## Declarations

### Ethics approval and consent to participate

This study involves computational experiments that are non-invasive and do not directly intervene with any human or animal subjects. Therefore, ethical approval from an institutional review board is not required. The consent of participants and the protection of personal information are not applicable, as the experimental data are sourced from publicly available datasets.

### Consent for publication

All authors have provided their consent for publication of this study. There are no identifiable individuals or personal data included in this manuscript, ensuring compliance with publication ethics.

### Competing interests

The authors declare no competing interests.

Received: 17 June 2024 Accepted: 1 October 2024

Published online: 08 October 2024

## References

- Fletcher EP, Madabushi R, Sahajwalla CG, Lesko LJ, Huang SM. The role of the FDA in guiding drug development. In: Atkinson's Principles of Clinical Pharmacology. 4th ed. Philadelphia: Elsevier; 2022. p. 681–90.
- Maharao N, Antontsev VG, Wright M, Varshney J. Entering the era of computationally driven drug development. *Drug Metab Rev*. 2020;52:283–98.
- Dai Y, Ge J. Clinical use of aspirin in treatment and prevention of cardiovascular disease. *Thrombosis*. 2012;2012(1):245037.
- Terentes-Printzios D, loakeimidis N, Rokkas K, Vlachopoulos C. Interactions between erectile dysfunction, cardiovascular disease and cardiovascular drugs. *Nat Rev Cardiol*. 2022;19(1):59–74.
- Samborsky O, Slobodyanyuk M, Malyi V. Strategic and tactful procedures of efficiency and investment attractiveness of new medicines development. *Chem Sci*. 2017;10:45.
- Hodos RA, Kidd BA, Shameer K, Readhead BP, Dudley JT. In silico methods for drug repurposing and pharmacology. *Wiley Interdiscip Rev Syst Biol Med*. 2016;8(3):186–210.
- Zhao BW, Su XR, Hu PW, Ma YP, Zhou X, Hu L. A geometric deep learning framework for drug repositioning over heterogeneous information networks. *Brief Bioinform*. 2022;23(6):bbac384.
- Zhao BW, Wang L, Hu PW, Wong L, Su XR, Wang BQ, et al. Fusing higher and lower-order biological information for drug repositioning via graph representation learning. *IEEE Trans Emerg Top Comput*. 2023;12(1):163–76.
- Su X, Hu P, Yi H, You Z, Hu L. Predicting drug-target interactions over heterogeneous information network. *IEEE J Biomed Health Inform*. 2022;27(1):562–72.
- Xie Z, Evangelopoulos X, Omar ÖH, Troisi A, Cooper AI, Chen L. Fine-tuning GPT-3 for machine learning electronic and functional properties of organic molecules. *Chem Sci*. 2024;15(2):500–10.
- Gallifant J, Fiske A, Levites Strelakova YA, Osorio-Valencia JS, Parke R, Mwauu R, et al. Peer review of GPT-4 technical report and systems card. *PLOS Digit Health*. 2024;3(1):e0000417.
- Boyko J, Cohen J, Fox N, Veiga MH, Li JI, Liu J, et al. An interdisciplinary outlook on large language models for scientific research. 2023. Available from: <https://arxiv.org/abs/2311.04929>.
- Hou W, Ji Z. GeneTuring tests GPT models in genomics. 2023. Available from: <https://www.biorxiv.org/content/10.1101/2023.03.11.532238v1>.
- Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, et al. Large language models generate functional protein sequences across diverse families. *Nat Biotechnol*. 2023;41(8):1099–106.
- Tang B, Tian N, Jiang J, Li Y, Yang J, Zhu Q. Investigation of zinc storage capacity of WS2 nanosheets for rechargeable aqueous Zn-ion batteries. *J Alloys Compd*. 2022;894:162391.
- Wang R, Feng H, Wei GW. ChatGPT in drug discovery: a case study on Anticocaine addiction drug development with Chatbots. *J Chem Inf Model*. 2023;63(22):7189–209.
- Sharma G, Thakur A. ChatGPT in Drug Discovery. *Chemistry*; 2023. Available from: <https://chemrxiv.org/engage/chemrxiv/article-details/63d56c13ae221ab9b240932f>
- Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. PMC-LLaMA: toward building open-source language models for medicine. *J Am Med Inform Assoc*. 2024;31:ocae045.
- Xie Q, Chen Q, Chen A, Peng C, Hu Y, Lin F, et al. Me LLaMA: foundation large language models for medical applications. 2024. Available from: <https://arxiv.org/abs/2402.12749>.
- Xue J, Wang YC, Wei C, Liu X, Woo J, Kuo CCJ. Bias and fairness in chatbots: an overview. 2023. Available from: <https://arxiv.org/abs/2309.08836>.
- Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877–901.
- ProbeMedicalYonseiMAILab. Medllama3-V20. 2024. Available from: <https://huggingface.co/ProbeMedicalYonseiMAILab/medllama3-v20>
- Zheng C, Liu Z, Xie E, Li Z, Li Y. Progressive-hint prompting improves reasoning in large language models. 2023. Available from: <https://arxiv.org/abs/2304.09797>.
- Gao H, Lin TE, Li H, Yang M, Wu Y, Ma W, et al. Self-explanation prompting improves dialogue understanding in large language models. 2023. Available from: <https://arxiv.org/abs/2309.12940>.
- Ye X, Iyer S, Celikyilmaz A, Stoyanov V, Durrett G, Pasunuru R. Complementary explanations for effective in-context learning. 2023. Available from: <https://arxiv.org/abs/2211.13892>.
- Diao S, Wang P, Lin Y, Pan R, Liu X, Zhang T. Active prompting with chain-of-thought for large language models. 2024. Available from: <https://arxiv.org/abs/2302.12246>.
- Arora S, Narayan A, Chen MF, Orr L, Guha N, Bhatia K, et al. Ask Me Anything: a simple strategy for prompting language models. 2023. Available from: <https://openreview.net/forum?id=bhUPjnS2g0X>.
- Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, et al. The RCSB protein data bank: redesigned web site and web services. *Nucleic Acids Res*. 2010;39(suppl\_1):D392–401.
- Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*. 2008;24(13):i232–40.
- Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multi-threading. *J Comput Chem*. 2010;31(2):455–61.
- Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev*. 2012;64:4–17.

32. Fu L, Shi S, Yi J, Wang N, He Y, Wu Z, et al. ADMETlab 3.0: an updated comprehensive online ADMET prediction platform enhanced with broader coverage, improved performance, API functionality and decision support. *Nucleic Acids Res.* 2024;52:gkae236.
33. Yang L, Chen H, Li Z, Ding X, Wu X. Give Us the Facts: enhancing large language models with knowledge graphs for fact-aware language modeling. 2024. Available from: <https://arxiv.org/abs/2306.11489>.
34. Bai P, Miljković F, John B, Lu H. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nat Mach Intell.* 2023;5(2):126–36.
35. Chen L, Fan Z, Chang J, Yang R, Hou H, Guo H, et al. Sequence-based drug design as a concept in computational drug design. *Nat Commun.* 2023;14(1):4217.
36. Harp JM, Hanson BL, Timm DE, Bunick GJ. Asymmetries in the nucleosome core particle at 2.5 Å resolution. *Acta Crystallogr D Biol Crystallogr.* 2000;56(12):1513–34.
37. Tachiwana H, Kagawa W, Osakabe A, Kawaguchi K, Shiga T, Hayashi-Takanaka Y, et al. Structural basis of instability of the nucleosome containing a testis-specific histone variant, human H3T. *Proc Natl Acad Sci.* 2010;107(23):10454–9.
38. Chakravarthy S, Gundimella SKY, Caron C, Perche PY, Pehrson JR, Khochbin S, et al. Structural characterization of the histone variant macroH2A. *Mol Cell Biol.* 2005;25:7616.
39. Wu B, Mohideen K, Vasudevan D, Davey CA. Structural insight into the sequence dependence of nucleosome positioning. *Structure.* 2010;18(4):528–36.
40. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. GROMACS: fast, flexible, and free. *J Comput Chem.* 2005;26(16):1701–18.
41. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M. PLIP: fully automated protein–ligand interaction profiler. *Nucleic Acids Res.* 2015;43(W1):W443–7.
42. Saberi Fathi SM, Tuszynski JA. A simple method for finding a protein’s ligand-binding pockets. *BMC Struct Biol.* 2014;14:1–9.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.